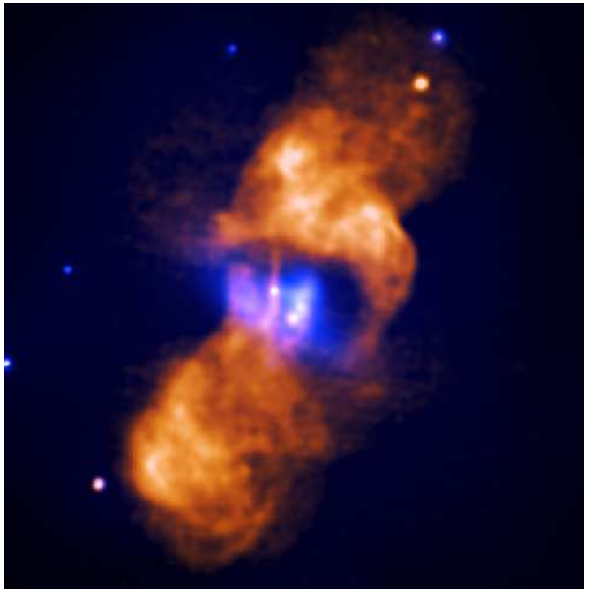
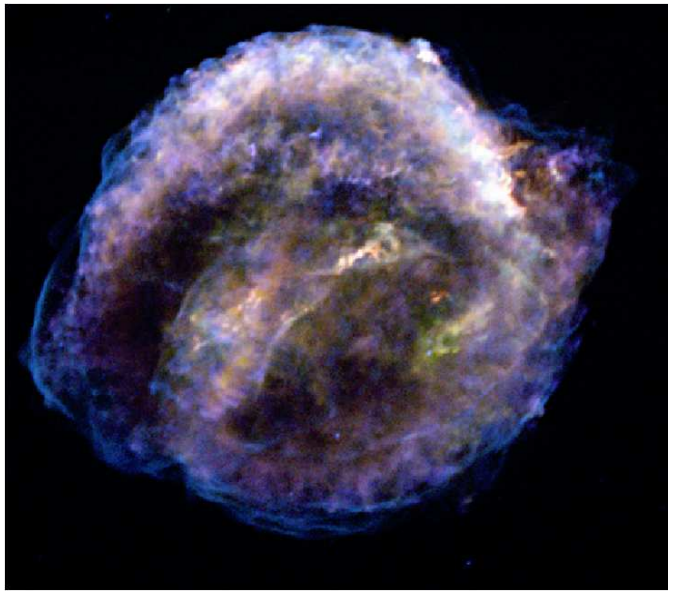
Partially Collapsed Gibbs Samplers
With Applications in High-Energy Astrophysics

David A. van Dyk

Department of Statistics
University of California, Irvine

Joint work with
Taeyoung Park





Outline of Presentation

A. **An idea from EM-type algorithms:**

Improve Computational Performance by Reducing Conditioning.

B. Applying this idea to Gibbs sampling: two simple examples.

C. An example from high-energy Astrophysics.

1. The Astronomical problem
2. Why the Gibbs sampler fails
3. Designing a new sampler

D. The general strategy and why it works.

The EM Algorithm

- The EM algorithm is a stable computation method designed to compute $\hat{\theta}$, the value of θ that optimizes

$$p(\theta|Y) = \int p(\theta, \psi|Y) d\psi.$$

- Here we embed the posterior distribution into a density on a larger space.

The EM Algorithm

E-STEP: Given $\theta^{(t)}$ compute

$$Q(\theta|\theta^{(t)}) = \mathrm{E}\left[\log p(\theta|Y, \psi) \mid Y, \theta^{(t)}\right].$$

M-STEP: Set data via the M-step

$$\theta^{(t+1)} = \mathrm{argmax}_\theta \; Q(\theta|\theta^{(t)})$$

The Corresponding Gibbs Sampler

STEP 1: Sample $\psi \sim p(\psi \mid Y, \theta^{(t)})$

STEP 2: Sample $\theta^{(t+1)} \sim p(\theta \mid Y, \psi)$

It is well known that the algorithms exhibit similar convergence behavior.

Improving the Rate of Convergence of EM

The EM Matrix Rate of Convergence

- EM is approximately an liner iteration, $(\theta^{(t+1)} - \hat{\theta}) \approx DM(\theta^{(t)} - \hat{\theta})$, with matrix rate of convergence

$$DM = I - I_{\text{obs}} I_{\text{aug}}^{-1},$$

where

$$I_{\text{obs}} = \frac{\partial^2}{\partial\theta \cdot \partial\theta} \log p(\theta \mid Y)\Big|_{\theta=\hat{\theta}} \quad \text{and} \quad I_{\text{aug}} = \frac{\partial^2}{\partial\theta \cdot \partial\theta} \log Q(\theta \mid \hat{\theta})\Big|_{\theta=\hat{\theta}}$$

- Meng and van Dyk (1996) showed that replacing ϕ with $g(\phi)$ can only improve the rate of convergence of EM.
- A similar strategy in the context of the corresponding Gibbs Sampler would construct a sampler on a smaller space.

The ECM, ECME, and AECM Algorithms

- We can replace STEP 2 of this Gibbs sampler with K draws:

$$\theta_k \sim p(\theta_k | Y, \theta_{-k}, \psi)$$

- Likewise the ECM algorithm replace the M-STEP of EM by a series of K conditional maximization or CM-STEPS:

$$\theta^{(t+\frac{k}{K})} = \text{argmax E}\left[\log p(\theta|Y,\psi) \mid Y, \theta^{(t)}\right] \text{ with } \theta_j^{(t+\frac{k}{K})} = \theta_j^{(t+\frac{k-1}{K})} \text{ for } j \neq k.$$

- To improve the rate of convergence of ECM, Liu and Rubin (1995) suggested replacing one or more of these CM-STEPS with:

$$\theta^{(t+\frac{k}{K})} = \text{argmax } \log p(\theta|Y) \text{ with } \theta_j^{(t+\frac{k}{K})} = \theta_j^{(t+\frac{k-1}{K})} \text{ for } j \neq k.$$

- In the more general framework of AECM, Meng and van Dyk (1997) replaced ψ with some function $g_k(\psi)$ in each of the CM-steps of ECM.

There is no Gibbs-like analogy to either ECME or AECM.

The Basic Idea

Replacing ψ with $g_k(\psi)$ in the CM-STEP,

$$\theta^{(t+\frac{k}{K})} = \text{argmax E}\left[\log p(\theta|Y,\psi) \mid Y,\theta^{(t)}\right] \text{ with } \theta_j^{(t+\frac{k}{K})} = \theta_j^{(t+\frac{k-1}{K})} \text{ for } j \neq k.$$

improves the computational performance by reducing conditioning.

Big Questions:

1. Can we employ a similar idea in the framework of Gibbs?

2. What happens if we replace the draw

$$\theta_k \sim p(\theta_k|Y,\psi,\theta_{-k}) \text{ with } \theta \sim p(\theta_k|Y,g(\psi),\theta_{-k})?$$

Meng and van Dyk (1997) showed that the order of the steps of ECME and AECM can effect the celebrated monotone convergence of EM-type algorithms.

3. Is there a similar effect in the analogous Gibbs-type samplers?

Our Goal is to Answer these Questions.

The Basic Idea

Replacing ψ with $g_k(\psi)$ in the CM-STEP,

$$\theta^{(t+\frac{k}{K})} = \text{argmax E}\left[\log p(\theta|Y,\psi) \mid Y,\theta^{(t)}\right] \text{ with } \theta_j^{(t+\frac{k}{K})} = \theta_j^{(t+\frac{k-1}{K})} \text{ for } j \neq k.$$

improves the computational performance by reducing conditioning.

Big Questions:

1. Can we employ a similar idea in the framework of Gibbs?

2. What happens if we replace the draw

$$\theta_k \sim p(\theta_k|Y,\psi,\theta_{-k}) \text{ with } \theta \sim p(\theta_k|Y,g(\psi),h(\theta_{-k}))?$$

Meng and van Dyk (1997) showed that the order of the steps of ECME and AECM can effect the celebrated monotone convergence of EM-type algorithms.

3. Is there a similar effect in the analogous Gibbs-type samplers?

Our Goal is to Answer these Questions.

Outline of Presentation

A. An idea from EM-type algorithms:

Improve Computational Performance by Reducing Conditioning.

B. **Applying this idea to Gibbs: two simple examples.**

C. An example from high-energy Astrophysics.

1. The Astronomical problem
2. Why the Gibbs sampler fails
3. Designing a new sampler

D. The general strategy and why it works.

Reducing Conditioning in Gibbs: The Simplest Example

Consider a simple two-step Gibbs sampler:

$$\begin{array}{ll} \textsc{Step 1:}\ \psi^{(t+1)} \sim p(\psi|\theta^{(t)}) & \\ \textsc{Step 2:}\ \theta^{(t+1)} \sim p(\theta|\psi^{(t+1)}) & \end{array} \quad \Longrightarrow \quad \begin{array}{l} \textsc{Step 1:}\ \psi^{(t+1)} \sim p(\psi|\theta^{(t)}) \\ \textsc{Step 2:}\ \theta^{(t+1)} \sim p(\theta), \end{array}$$

where we replace $\psi^{(t+1)}$ with $g(\psi^{(t+1)}) = c$ in Step 2.

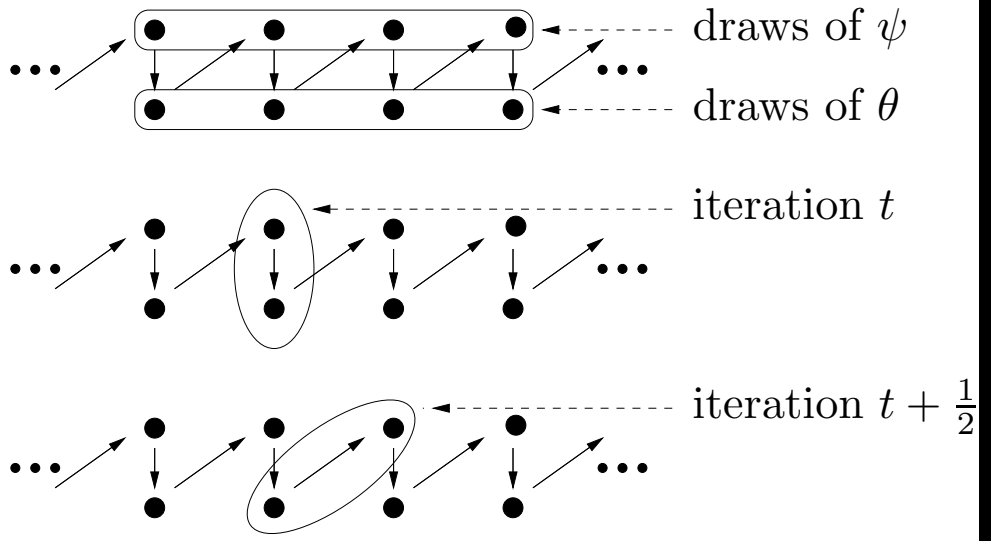
The Markov chain

$$\mathcal{M} = \{(\psi^{(t)}, \theta^{(t)}), t = 0, 1, \ldots\}$$

has stationary dist'n $p(\psi)p(\theta)$

- with target margins but
- without the correlation of the target distribution,

AND converges quickly!



We regain the joint target distribution with a one-step shifted chain.

Heads Up!

Reducing the conditioning within Gibbs-type samplers involves new challenges:

- The order of the draws may effect the stationary distribution of the chain.
- The conditional distributions may not be compatible with *any* joint distribution.
- The steps sometimes can be blocked to form an ordinary Gibbs sampler with fewer steps.

A Surrogate Distribution

- The joint distribution of our modified simple Gibbs sampler is

$$p(\psi)p(\theta)$$

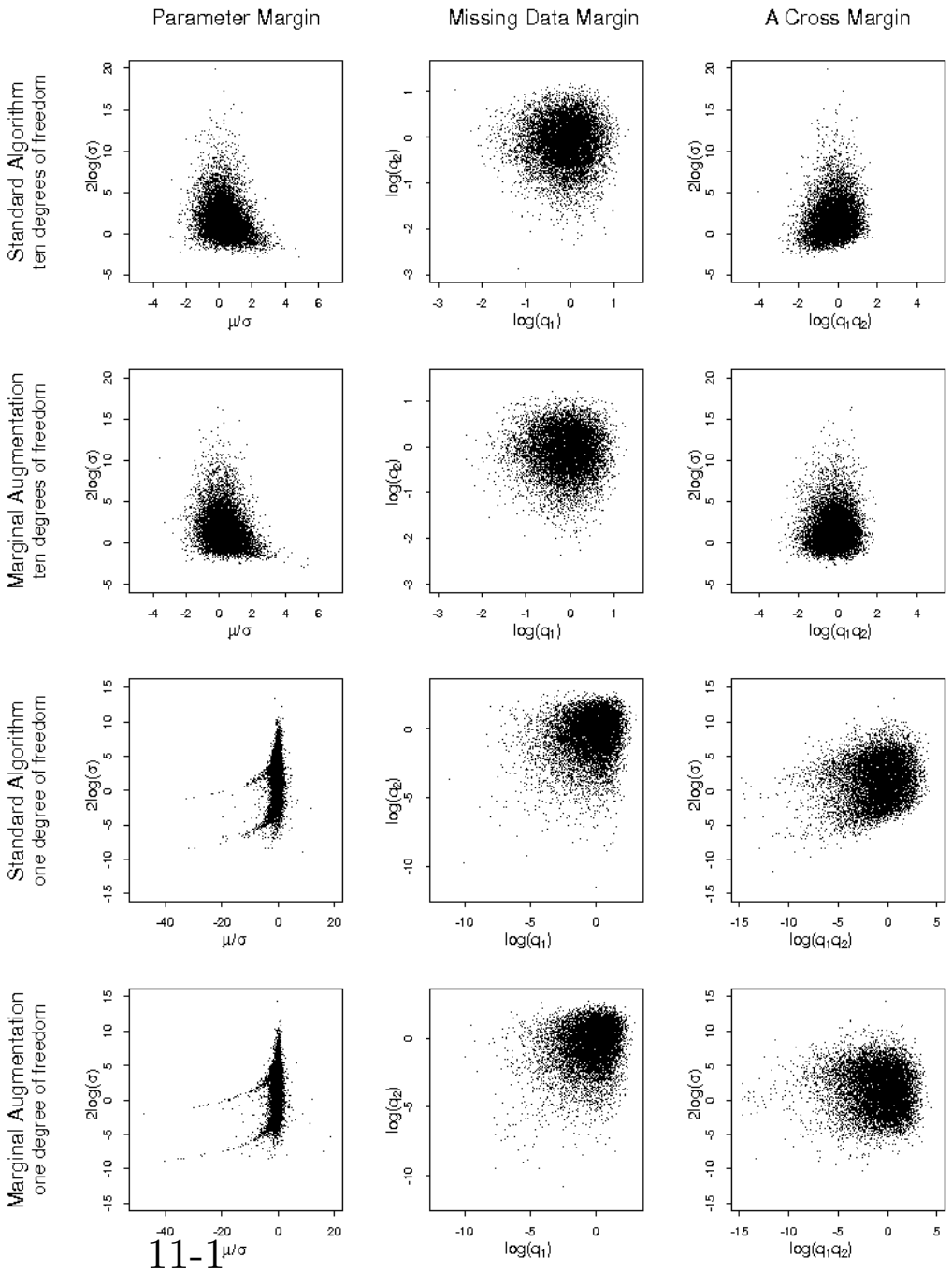
  rather than

$$p(\psi, \theta).$$

- These are two joint distributions with the same marginal distributions but with different correlation structures.
- By taking one conditional distribution from each we create a chain with
  1. a stationary distribution that oscillates between the two joint distributions,
  2. reduced autocorrelation due to the lower correlation in the *surrogate* joint distribution, and
  3. a stationary distribution for the marginal chains that is equal to the corresponding margin of the target distribution.

Empirical Illustration with a t model

- *The loss of the correlation structure is our key to success.*
- Two 'data sets' of size two are fit with 10 and 2 degrees of freedom.
- These algorithms are based on the method of *Marginal Augmentation* (Meng and van Dyk, 1999; van Dyk and Meng, 2001).
- Idea: Use both conditionals of the joint distribution with reduced correlation.



Outline of Presentation

A. An idea from EM-type algorithms:

Improve Computational Performance by Reducing Conditioning.

B. Applying this idea to Gibbs sampling: two simple examples.

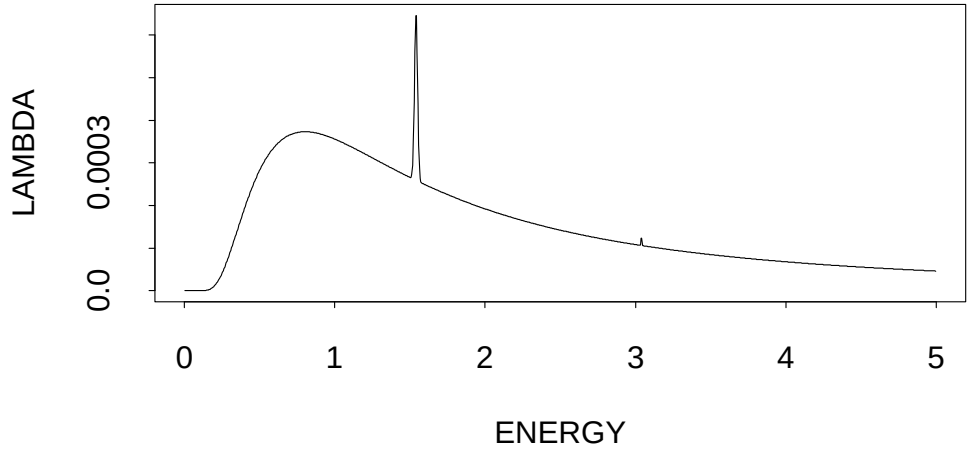
C. **An example from high-energy Astrophysics.**

1. **The Astronomical problem**
2. **Why the Gibbs sampler fails**
3. **Designing a new sampler**

D. The general strategy and why it works.

The Basic Spectral Models

- Spectral Analysis: Modeling the distribution of photon energies.
- Data: Counts in narrow energy bins.
- Counts modeled with Poisson process, with parameter varying with energy.
  1. The *continuum,* a GLM for the baseline spectrum (e.g., $\alpha E^{-\beta}$),
  2. Several *emission lines*, a mixture of Gaussians added to the continuum.
  3. Several *absorption features* multiply by the continuum.
  4. The continuum indicates the temperature of the source while the emission and absorption lines gives clues as to the relative abundances of elements.



Searching for Narrow Lines

- A simplified *latent* Poisson Process for the scientific model,

$$X_i \sim \text{Poisson}\left(\Lambda_i = \alpha E_i^{-\beta} + \lambda^{\text{L}} \pi_i\right).$$

- We sometimes construct a *delta function* emission line model so that
  1. the emission line is contained entirely in one bin, but
  2. we do not know which bin.

  I.e., $\{\pi_i\}$ can be parameterized in terms of a single unknown parameter,

$$\theta^{\text{L}} = \text{the location of the emission line.}$$

- Using *Data Augmentation* to fit this finite mixture model:

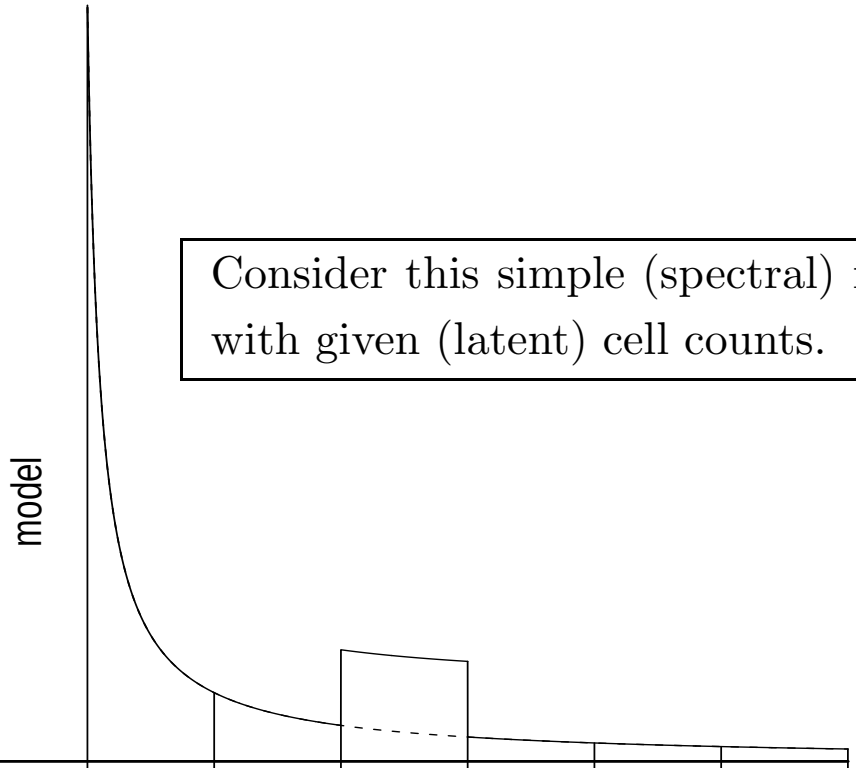
$$Z_{il} = \begin{pmatrix}\text{indicator that photon } l \text{ in bin } i \\ \text{corresponds to the emission line}\end{pmatrix}$$

  1. Given $Z = \{Z_{il}\}$ we can sample $\theta = \{\alpha, \beta, \lambda^{\text{L}}, \theta^{\text{L}}\}$
  2. Given θ we can sample Z, via $Z_{il} \sim \text{Ber}\left(\frac{\lambda^{\text{L}} \pi_i}{\alpha E_i^{-\beta} + \lambda^{\text{L}} \pi_i}\right)$

In This Case the Gibbs Sampler Fails.

Why the Gibbs Sampler Fails

Consider this simple (spectral) model with given (latent) cell counts.

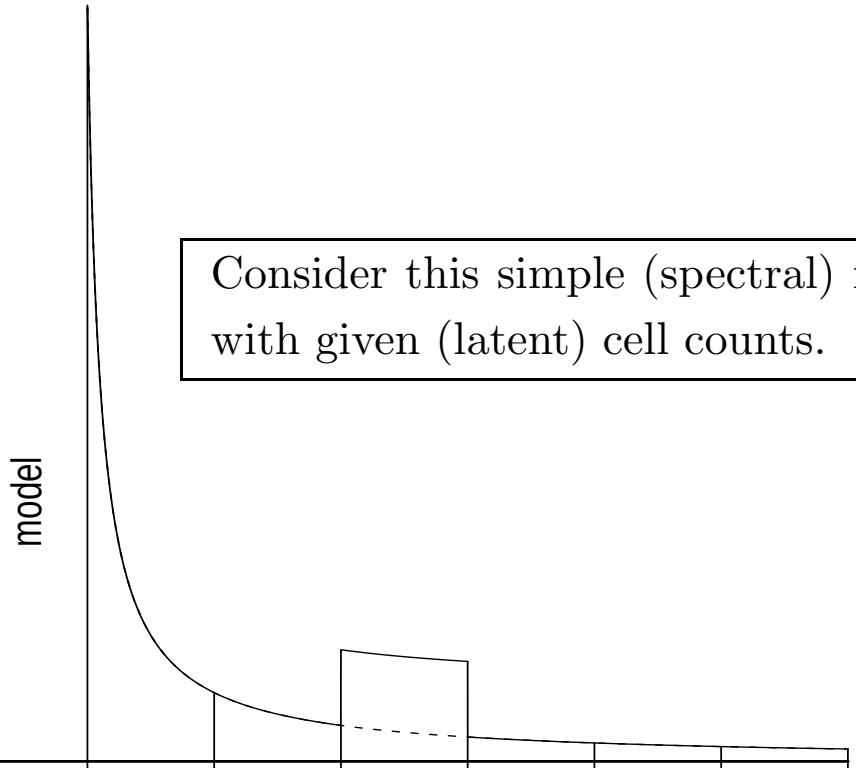


X = (latent) Cell Counts	10	4	8	1	2	0
Continuum Counts(Z=0)						
Line Counts (Z=1)						

Given this Model, what is Z?

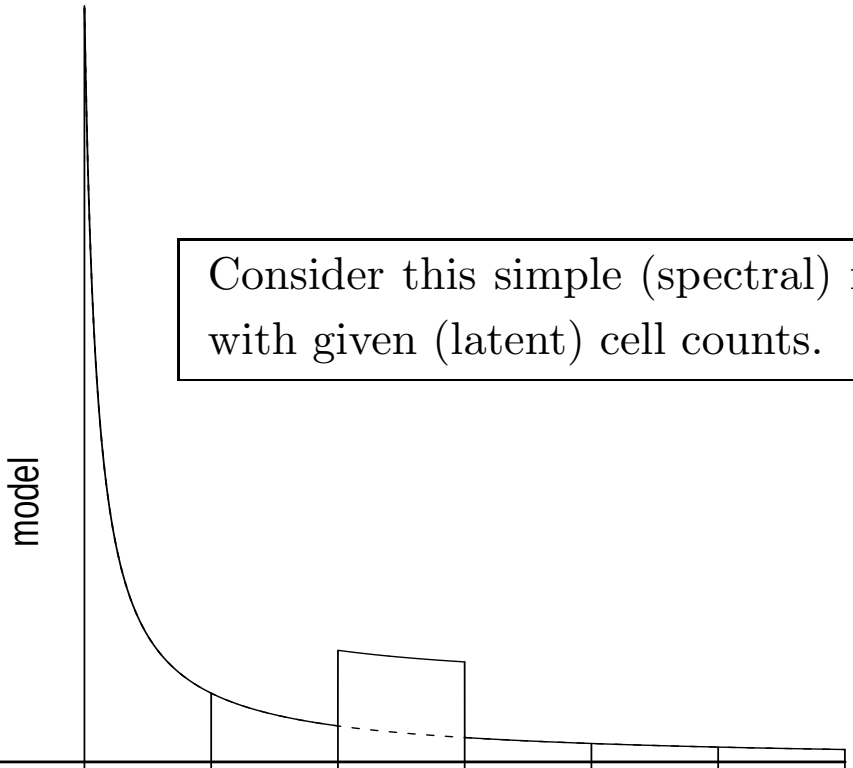
Why the Gibbs Sampler Fails

Consider this simple (spectral) model with given (latent) cell counts.



X = (latent) Cell Counts	10	4	8	1	2	0
Continuum Counts(Z=0)	10	4	~3	1	2	0
Line Counts (Z=1)	0	0	~5	0	0	0

Why the Gibbs Sampler Fails



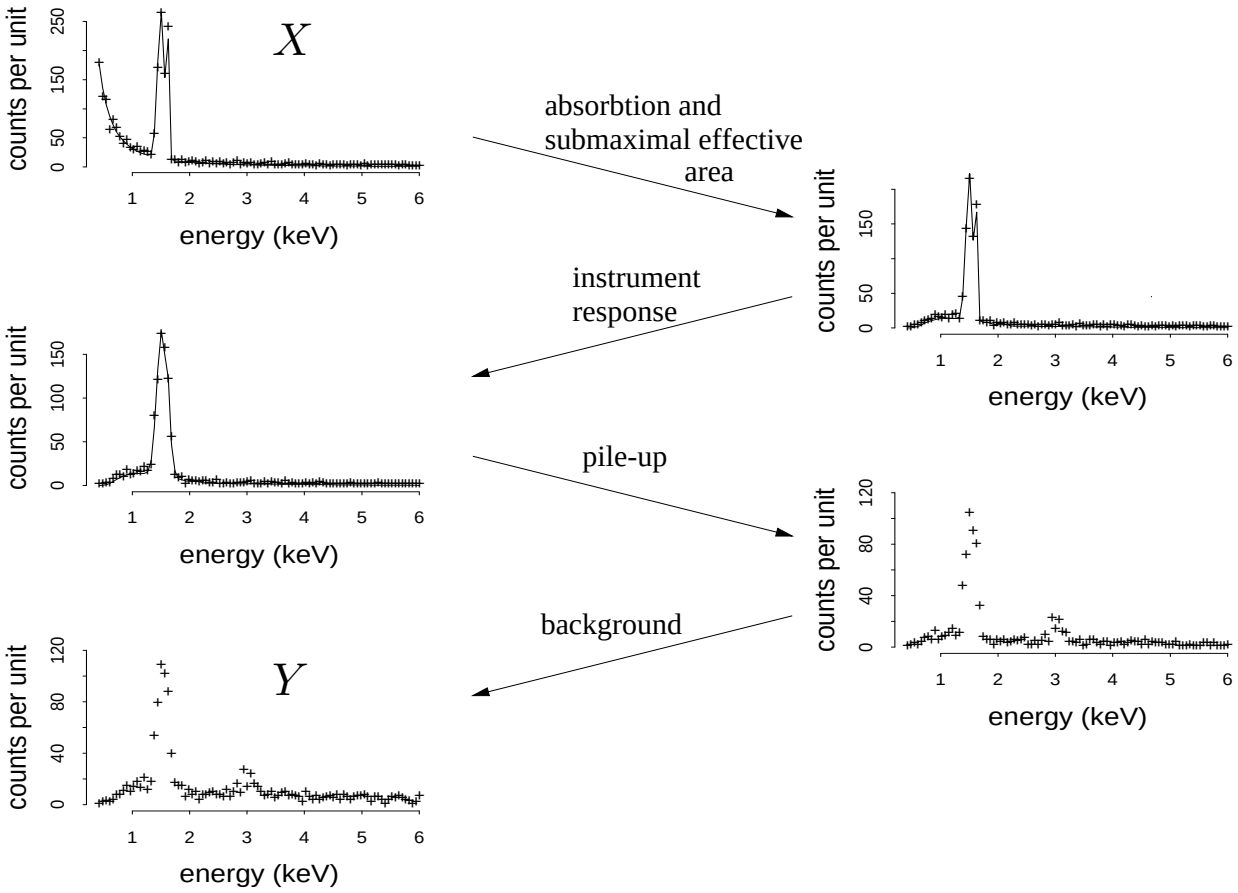
Consider this simple (spectral) model with given (latent) cell counts.

X = (latent) Cell Counts	10	4	8	1	2	0
Continuum Counts(Z=0)	10	4	~3	1	2	0
Line Counts (Z=1)	0	0	~5	0	0	0

Given Z, what is the location of the emission line?

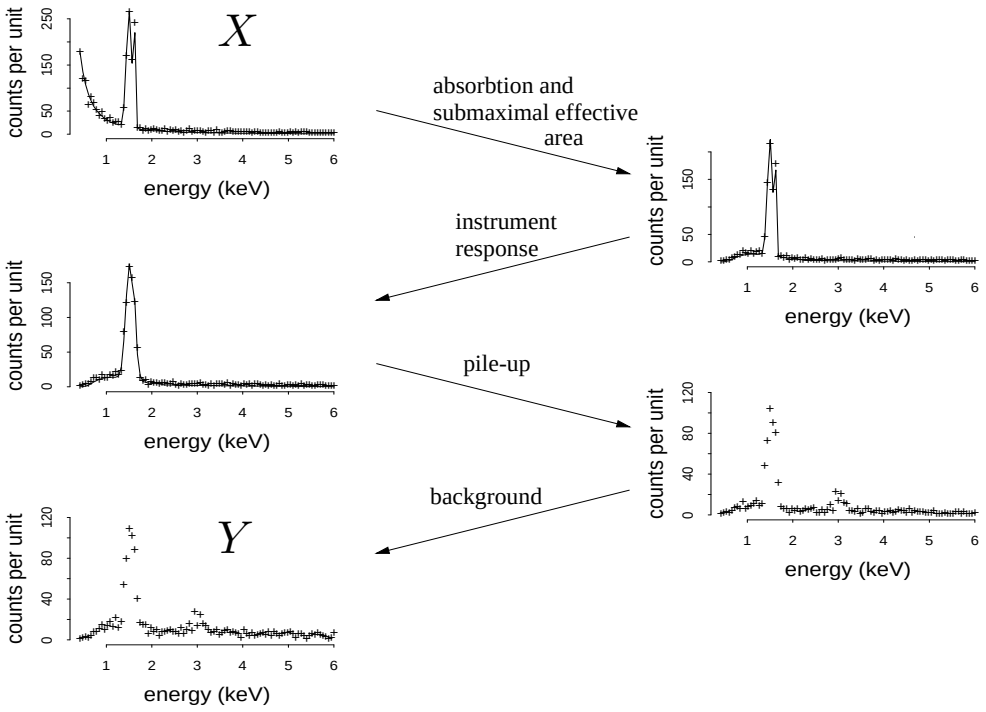
Highly Structured Models

This model is a simplification. To see how the actual samplers work, we need to model the data collection mechanism.



Highly Structured Models

Modeling the *Chandra* data collection mechanism.



- The method of Data Augmentation: EM algorithms and Gibbs samplers.
- We can separate a complex problem into a sequence of problems, each of which is easy to solve.

We wish to directly model the sources and data collection mechanism and use statistical procedures to fit the resulting highly-structured models and address the substantive scientific questions.

The Standard Gibbs Sampler

We do not observe the latent Poisson Process,

$$X_i \sim \text{Poisson}\left(\Lambda_i = \alpha E_i^{-\beta} + \lambda^{\text{L}}\pi_i\right),$$

Rather we observe, $Y_j \sim \text{Poisson}\left(a_j \sum_i P_{ij}\Lambda_i + \xi_j\right)$

Y_{obs}	=	$\{Y_j\} =$ obs cell cnts
X	=	$\{X_i\} =$ latent cell cnts
Z	=	emission line indicators
θ^{L}	=	location of emission line
θ^{O}	=	other model parameters

The standard Gibbs sampler simulates:

1. $p(X, Z|\theta)$
2. $p(\theta|X, Z) = p(\theta^{\text{O}}|X, Z)p(\theta^{\text{L}}|X, Z)$

We tacitly condition on Y_{obs} throughout.

With a delta function emission line model, this sampler fails.

An Incompatible Gibbs Sampler

- Recall the "Simplest Example":

$$\begin{array}{l} p(\psi|\theta) \\ p(\theta|\psi) \end{array} \longrightarrow \begin{array}{l} p(\psi|\theta) \\ p(\theta) \end{array} \longrightarrow \begin{array}{l} p(\theta) \\ p(\psi|\theta) \end{array} \longrightarrow p(\theta, \psi)$$

- Following this we construct:

  Sampler 1: (A Blocked Version of the Original Sampler.)

$$\begin{array}{l} p(X, Z|\theta) \\ p(\theta^{\mathrm{O}}|\theta^{\mathrm{L}}, X, Z) \\ p(\theta^{\mathrm{L}}|\theta^{\mathrm{O}}, X, Z) \end{array} \longrightarrow \begin{array}{l} p(X, Z|\theta) \\ p(\theta^{\mathrm{O}}|\theta^{\mathrm{L}}, X, Z) \\ p(\theta^{\mathrm{L}}|\theta^{\mathrm{O}}) \end{array} \longrightarrow \begin{array}{l} p(\theta^{\mathrm{L}}|\theta^{\mathrm{O}}) \\ p(X, Z|\theta) \\ p(\theta^{\mathrm{O}}|\theta^{\mathrm{L}}, X, Z) \end{array} \longrightarrow \begin{array}{l} p(\theta^{\mathrm{L}}, X, Z|\theta^{\mathrm{O}}) \\ p(\theta^{\mathrm{O}}|\theta^{\mathrm{L}}, X, Z) \end{array}$$

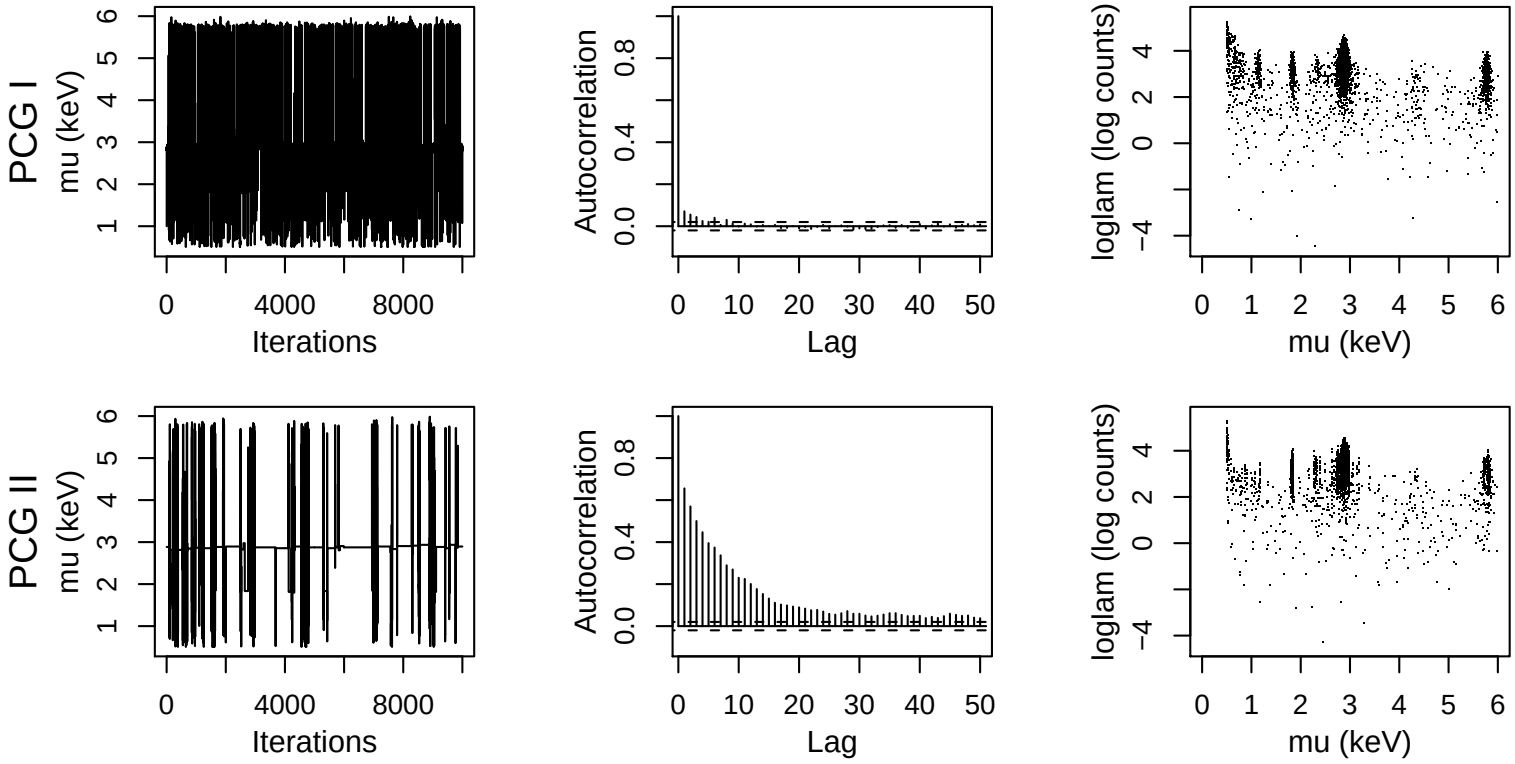
  Sampler 2: (Cannot be Blocked: An Incompatible Gibbs Sampler.)

$$\begin{array}{l} p(X, Z|\theta) \\ p(\theta^{\mathrm{O}}|\theta^{\mathrm{L}}, X, Z) \\ p(\theta^{\mathrm{L}}|\theta^{\mathrm{O}}, X, Z) \end{array} \longrightarrow \begin{array}{l} p(X, Z|\theta) \\ p(\theta^{\mathrm{O}}|\theta^{\mathrm{L}}, X, Z) \\ p(\theta^{\mathrm{L}}|\theta^{\mathrm{O}}, X) \end{array} \longrightarrow \begin{array}{l} p(\theta^{\mathrm{L}}|\theta^{\mathrm{O}}, X) \\ p(X, Z|\theta) \\ p(\theta^{\mathrm{O}}|\theta^{\mathrm{L}}, X, Z) \end{array}$$

It can be shown that both samplers have the correct stationary distribution and are faster to converge than the standard sampler.

Computational Gains

- Compare Standard Sampler, Sampler 1, and Sampler 2 in a spectral analysis.
- Standard sampler doesn't move from its starting value.
- Sampler 1 has much better convergence characteristics than Sampler 2.
- However, each iteration of Sampler 1 is more expensive.



Verifying the Stationary Distribution of Sampler 2

$$
\begin{array}{lcl}
p(X, Z|\theta) & & p(X, Z|\theta) \\
p(\theta^{\mathrm{O}}|\theta^{\mathrm{L}}, X, Z) & \longrightarrow & p(\theta^{\mathrm{O}}|\theta^{\mathrm{L}}, X, Z) \\
p(\theta^{\mathrm{L}}|\theta^{\mathrm{O}}, X, Z) & & p(\theta^{\mathrm{L}}, Z|\theta^{\mathrm{O}}, X)
\end{array}
$$

We move Z to the left of the conditioning sign in Step 3.

$$
\begin{array}{lcl}
 & & p(\theta^{\mathrm{L}}, Z|\theta^{\mathrm{O}}, X) \\
 & \longrightarrow & p(X, Z|\theta) \\
 & & p(\theta^{\mathrm{O}}|\theta^{\mathrm{L}}, X, Z)
\end{array}
$$

We permute the order of the steps.

$$
\begin{array}{lcl}
 & & p(\theta^{\mathrm{L}}|\theta^{\mathrm{O}}, X) \\
 & \longrightarrow & p(X, Z|\theta) \\
 & & p(\theta^{\mathrm{O}}|\theta^{\mathrm{L}}, X, Z)
\end{array}
$$

We remove Z from the draw in Step 1, since the transition kernel does not depend on this quantity.

None of these operations effect the chain's stationary distribution.

Outline of Presentation

A. An idea from EM-type algorithms:

   Improve Computational Performance by Reducing Conditioning.

B. Applying this idea to Gibbs sampling: two simple examples.

C. An example from high-energy Astrophysics.

   1. The Astronomical problem
   2. Why the Gibbs sampler fails
   3. Designing a new sampler

D. **The general strategy and why it works.**

The General Strategy

1. Marginalzing

$$\begin{array}{lcl} p(X, Z|\theta) & & p(X, Z|\theta) \\ p(\theta^{\mathrm{O}}|\theta^{\mathrm{L}}, X, Z) & \longrightarrow & p(\theta^{\mathrm{O}}|\theta^{\mathrm{L}}, X, Z) \\ p(\theta^{\mathrm{L}}|\theta^{\mathrm{O}}, X, Z) & & p(\theta^{\mathrm{L}}, Z|\theta^{\mathrm{O}}, X) \end{array}$$

Move quantities from the right to the left of the conditioning sign. This does not alter the stationary distribution, but improves the rate of convergence.

2. Permuting

$$\begin{array}{lcl} & & p(\theta^{\mathrm{L}}, Z|\theta^{\mathrm{O}}, X) \\ & \longrightarrow & p(X, Z|\theta) \\ & & p(\theta^{\mathrm{O}}|\theta^{\mathrm{L}}, X, Z) \end{array}$$

Permute the order of the steps. This can have minor effects on the rate of convergence, but does not affect the stationary distribution.

3. Trimming

$$\begin{array}{lcl} & & p(\theta^{\mathrm{L}}|\theta^{\mathrm{O}}, X) \\ & \longrightarrow & p(X, Z|\theta) \\ & & p(\theta^{\mathrm{O}}|\theta^{\mathrm{L}}, X, Z) \end{array}$$

Remove quantities that are not part of the transition kernel. This does not effect the stochastic mapping or the rate of convergence.

The Advantage of Partially Collapsing

An Outline of a proof:

- The dependence of consecutive iterations of the Gibbs Sampler flows through what is conditioned upon in the *first step* of each iteration.
- The maximal autocorrelation can only decrease if we reduce this conditioning. Compare $\mathcal{K}(\theta \mid \theta')$ with $\mathcal{K}(\theta \mid g(\theta'))$.
- The Spectral Radius of the Chain
  - generally governs convergence,
  - is bounded above by the maximal autocorrelation, and
  - does not depend on which step begins the iteration, as long as the order of steps is not altered.

By reducing conditioning in any step (i.e., partially collapsing) we reduce both a bound on the spectral radius of the chain and the maximal autocorrelation for the chain that starts with that step.

References

- Park, T. and van Dyk, D. A. (2009). Partially Collapsed Gibbs Samplers: Illustrations and Applications. *Journal of Computational and Graphical Statistics*, in press.
- van Dyk, D. A. and Meng, X. L. (2009). Cross-Fertilizing Strategies for Better EM Mountain Climbing and DA Field Exploration: A Graphical Guide Book. *Statistical Science*, under revision.
- Park, T., van Dyk, D. A., and Siemiginowska, A. (2008). Searching for Narrow Emission Lines in X-ray Spectra: Computation and Methods *The Astrophysical Journal*, **688**, 807-825.
- van Dyk, D. A. and Park, T. (2008). Partially Collapsed Gibbs Samplers: Theory and Methods. *Journal of the American Statistical Association*, **103**, 790–796.